DOCUMENT RESUME

ED 144 527 IR 004 775

AUTHOR Atherton, Pauline
TITLE Subject Access Project. Second Quarterly Progress Report.
INSTITUTION Syracuse Univ., N.Y. School of Information Studies.
PUB DATE Dec 76
NOTE 12p.; For related documents, see IR 004 620 and ED 129 229

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Cataloging; Content Analysis; Data Bases; *Experimental Programs; *Indexing; Information Retrieval; Library Research; On Line Systems; Subject Index Terms
IDENTIFIERS Machine Readable Cataloging; *Subject Access

ABSTRACT

The Subject Access Project second quarterly report for September to December 1976 summarized in-progress work to improve subject access to monographs. Activities include: (1) analysis of book indexes and tables of contents for terms to augment MARC subject description; (2) analysis of additional book sections--e.g., maps, illustrations, charts, and tables for headings coverage in indexes and contents pages; (3) examination of significant characteristics of book indexes; (4) preparation of a manual of rules and guidelines for selection of entries from table of contents or subject index to assure usability of terms in on-line searching; (5) calculation of a quota for maximum entries from each monograph; (6) preliminary cost analysis for data input; and (7) interview with prospective users to determine user requirments of the on-line catalog. The interview form is attached. (KP)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED144527

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SYRACUSE UNIVERISTY

SCHOOL OF INFORMATION STUDIES

SUBJECT ACCESS PROJECT

SECOND QUARTERLY PROGRESS REPORT

December, 1976

PROJECT STAFF

Pauline Atherton - Project Director
B.K.L. Genova - Assistant Project Director
Bette Brindle - Research Associate
Barbara Settel - Research Assistant
Cheryl Willson - Secretary
Bonnie Gratch - Part-time Research Assistant

Summary of Work in Progress

(1) Analysis of Sample -- Updated Table

(2) Study of Book Indexes for Subject Retrieval

(3) Analysis of Front Matter

(4) Choice of T/C or Index for Input; Selection Rules for Entries -- A Manual for Augmenting Subject Descriptions of Books

(5) How Many is Enough? A Formula for Maximum Number of Entries

(6) Preliminary Cost Analysis for Data Input

(7) Determining User Requirements

Appendix: A. Interview Guide for Determining User Requirements

B. Informal Financial Report

IR004775

(1) ANALYSIS OF SAMPLE -- UPDATED TABLE AND TAPE

We have completed processing all available items. Ninety-five items have been deleted from the original sample because they were duplicates, non-English books, journal articles erroneously listed as monographs, or unobtainable, which is an attrition rate of 4%. In addition, we have found that 253 items did not contain a table of contents or an index (approximately 11% of our current sample) and these will be deleted later.

Table One below reports the status of our monograph analysis as follows:

Column 1 - number of items located, processed, and ready for the next phase
Column 2 - indicates the number of items which did not contain a table of contents or index
Column 3 - number of processed items in Column 1 that have usable table of contents
Column 4 - percent of the processed that contain table of contents
Column 5 - number of processed (Column 1) items that have an index
Column 6 - percent of processed (Column 1) items that have an index
Column 7 - the range of the number of pages per index

On the basis of this analysis, the tape of m-r records we received from UTLAS will be updated and made ready for our input to the records. Also, we plan to prepare short-title lists for cooperating libraries to check their catalogs against our sample to determine if they would find it useful to search our data base when it is up on SDC/Orbit.

TABLE 1

Sample	# of Items (1)	Items Containing No T/C or Index (2)	Table of Contents #(3)	Table of Contents (4)	Index #(5)	Index (6)	Range of # of Index Pages (7)
1. Philosophy BC Logic BH Aesthetics BJ Ethics	151	9	142	94%	84	56%	1-74
2. History DE Graeco-Roman World DF Ancient Greece DG Ancient Italy Rome to 476	137	4	130	95	110	80	1-141
3. Arts NB Sculpture NE Engraving NK Art Applied to Industry Decoration & Ornament	196	31	140	71	92	47	2-94
4. Literature PN 1560-3300	311	31	269	87	162	52	1-35
5. Psychology BF 1-990	342	23	331	97	269	79	1-117
6. Anthropology GN 1-696	142	6	133	94	95	67	1-42
7. Public Finance HJ	133	18	114	86	77	58	1-76
8. Sociology HM 1-221	154	7	147	95	107	69	1-27
Sample Totals	1566	129	1406	90%	996	64%	1-141
Complete Collection Listings							
1. Urban Planning. Urban Redevelopment HT 166-177	521	35	480	92	151	29	1-43
2. Post Confederation Ontario History F 5520-5547	294	89	191	65	89	30	1-89
Complete Collection Total	815	124	671	82%	240	29%	1-89
Grand Total	2381	253	2077	87%	1236	52%	1-141

(2) STUDY OF BOOK INDEXES FOR SUBJECT RETRIEVAL

In the process of analyzing book indexes for the selection of terms to augment the subject description on the MARC records for books, we have collected a great deal of data on the characteristics of book indexes. To document our findings, we are preparing a paper which will:

(1) provide an empirical analysis of characteristics of book indexes in ten subject areas within the social sciences and the humanities;

(2) describe the present inadequacies of book indexes for subject retrieval; and

(3) make suggestions for the preparation of indexes, to improve their potential use in machine-readable bibliographic records.

This study is based on an examination of 123 books with indexes (a 10% sample from our collection of over 1200 books with subject indexes in ten subject areas).

Significant characteristics of book indexes are grouped into 8 categories and are quantified and evaluated by subject area. Categories of analysis are:

(1) length of index (averages of index pages, lines per page, and pages of books);

(2) arrangement of index (a single dictionary index, multiple arrangement, or single personal names or subject index);

(3) scope of index (explanation of scope, front or back matter indexed, maps or illustrations indexed);

(4) entry headings (number of personal names, subject terms and title headings, multiple word headings in inverted or natural language, headings predominantly generic or specific, existence of pre-coordination of terms in headings, consistency of type and style maintained);

(5) locators (type of locator, use of special face type, non-subheaded entries with 10 or more locators);

(6) subheadings (number of indexes that do and do not have subheaded entries; style and arrangement of subheadings);

(7) control devices (scope notes, type of "see" and "see also" references, multiple entries);

(8) physical appearance (margins, spacing and type size).

The choice of the above characteristics and the subsequent evaluation represent a synthesis from several sources: the "U.S.A. Standard Basic Criteria for Book Indexes" (ANSI Z39.4, revised 1968); the British Standards Institution "Recommendations for the Preparation of Indexes" (BS 3700: 1976); the University of Chicago's, "Manual of Style" (pp. 399-430); and two symposia on book indexes reported in

The Indexer ("The Inadequacies of Book Indexes," Vol. 9, No. 1, April 1974 and "Symposium on Selective Indexing," Vol. 9, No. 2, October 1974).

Each index in the sample of 123 was examined page by page so that total tallies were recorded for four characteristics. The other four characteristics were tabulated from the second page of each index in the study.

A complete report is now being prepared for publication, but we can abstract some of the highlights of our findings here. Although there are some differences between the findings for the social sciences and humanities, considered together as a total, the book indexes analyzed fail to meet many of the most important "Basic Criteria" recommended by the U.S. and British Standards. The most significant inadequacies are:

(1) the preponderance of personal name entries;
(2) lack of an introductory note to explain the scope of the index;
(3) lack of sufficient control devices to permit multiple points of entry into the index with the result of scattering entries for a given concept or topic; and
(4) an excessive quantity of non-subheaded entries with 10 or more locators.

Only the last inadequacy will affect our work on the Subject Access Project.

(3) ANALYSIS OF "FRONT MATTER" IN BOOKS

In the process of developing our selection rules for using entries from contents and index pages as subject descriptions for books, we have made a separate analysis of the characteristics of other front matter in books (e.g., lists of maps, charts, illustrations, figures, tables, plates and diagrams). Samples were taken to determine the extent to which the headings in these lists are covered by the words in contents pages or indexes. This information was tabulated by subject area and by type of front matter.

The first important observation we can make is that tables have the highest proportion of terms and phrases which are not covered by tables of contents or indexes. Of the eight subject areas studied, Public Finance (HJ), Urban Planning (HT), and Sociology (HM) exhibited this finding the most conclusively. Our analysis shows that the captions of tables in these areas are highly quantitative, including statistical and historical (dates) information. The captions of these tables appeared very similar to the phrasing of queries -- statements which are made to an on-line computer-based searching service.

On the basis of this analysis we will include directions in our selection rules for choosing those entries in lists of tables which are not covered by entries in the contents tables or indexes. Wherever appropriate, additional selection from other front matter may also be made.

(4) CHOICE OF T/C OR INDEX FOR INPUT; SELECTION RULES FOR ENTRIES --

A Manual for Augmenting Subject Descriptions

In preparation for the task of augmenting subject descriptions we have developed a manual of specific rules and guidelines for

(1) Choice of Table of Contents (T/C) over Subject Index (when both are available)

(2) Selection of Entries from T/C and/or Index (to assure usability of terms and phrases in on-line searching)

The staff has tested the manual for training others to perform these tasks. A preliminary edition of the manual is being typed and will be available for student helpers who will begin work during Christmas holiday period.

We also developed a quota for maximum number of entries to select (see next section of this report).

(5) HOW MANY IS ENOUGH?

A Formula for Maximum Number of Entries

As we have heard often, the number of subject headings on an LC catalog card is 1.3 regardless of size of book or subject area. In order to have a guideline for the number of entries we would select from the table of contents and/or index, we thought it best to estimate the maximum to be used for each monograph. It is expected that the use of this "quota" will allow us to reflect the major contents of the monograph. The process of calculating the quota is as follows:

(1) For each subject field, the average number of table of contents entries per monograph in the class and average number of possible index entries per monograph are computed.

(2) A ratio (R) of the two averages is then calculated:

$$\frac{\text{average index entries per monograph in class}}{\text{average table of contents entries per monograph in class}} = \text{Ratio,}$$

the relationship of the table of contents to the index.

(3) This R is then used as the divisor to calculate the quota for each monograph by dividing the number of index entries in it by R:

$$\frac{\text{Total number of index entries in monograph}}{R} = \text{"quota" of entries to be selected}$$

Example:

For (Anthropology) LC Class GN

average index entries per monograph = 862.79 entries
average table of contents entries = 61.88 entries

(1) Ratio: $\frac{\text{862.79 Index entries}}{\text{61.88 T/C entries}} = 14 = R$

(2) Quota for monograph with LC call number GN34.3.M3E95:

number of index entries = 416

$$\text{quota} = \frac{416}{14} = 30$$

(6) PRELIMINARY COST ANALYSIS FOR DATA INPUT

The entries to augment each MARC record subject description will be entered on-line via DEC terminals on to disk. A text editing program (SOS) on the DEC-10 system, rather than keypunching on to cards, will be used. The entries will be put on to tape after each set is verified and corrected for errors.

We estimate:

entry time per monograph -	29 minutes
time to correct errors -	5 minutes
	34 minutes

(costs: $3.00 per hour for staff time = $1.70 plus $.50/computer) = $2.20 per monograph.

(7) DETERMINING USER REQUIREMENTS

Before proceeding too fast toward a finalization of our selection rules, we decided to interview prospective users of such a new on-line catalog with augmented subject descriptions. Faculty and students at Syracuse University were approached and asked to give us queries to such a system (following certain "rules of the game"). We will continue to interview prospective users and will be aided in this effort by the Robarts Library Reference Department staff and librarians at three or four other academic institutions. An interview guide used for this study is given in the appendix to this report.

PLANS FOR NEXT QUARTER

(a) Train data selection and input staff.
(b) Produce half (or better) of data in machine-readable form.
(c) Ship some data to SDC for processing.
(d) Produce short title checklists of our data base.
(e) Run searches on BRS and SDC for practice.

SUBJECT ACCESS PROJECT

SYRACUSE UNIVERSITY AND UNIVERSITY OF TORONTO NORARTS LIBRARY

At the present time we think you cannot easily obtain references to information inside books (monographs). Library catalogs can only tell you what the book is generally about.

We want to change all that by augmenting the subject descriptions for books. Someday soon you should be able to type your query on a computer terminal and get an immediate response from the bibliographic retrieval service about the library's holdings on that subject.

You can help us determine how to improve the book's description in the library catalog by playing the following game. By your responses you will provide us with data that will help us re-design the record for books in the library's computer-based catalog.

Assume for the purposes here that:

(1) you are describing your request to a colleague in your field who would understand you without any difficulty.

(2) you can pass your request on to an "ideal" system you can trust to provide you with a list of references to parts of books pertinent to your request.

(3) you can specify how many pages of the book must be on your subject before it is retrieved for you.

(4) you can state your query in near "natural language" with no need to use a special list of subject headings or classification numbers.

(5) you can specify the time period to be covered or geographic restrictions (e.g., Public financial policy in Great Britain in 19th century).

(6) you can modify your request by restricting output to books published after a certain date or in a certain country of publication or language.

(7) if you have a particular orientation (e.g., main interest is psychological aspects, not sociological, etc.), you could specify this and it would help to refine the output to those books with a similar orientation.